Original article

Journal of Defense Modeling and
Simulation: Applications,
Methodology, Technology
XX(X) 1–11
© 2010 The Society for Modeling
and Simulation International
DOI: 10.1177/1548512910382002
http://dms.sagepub.com

Online adaptive learning for team strategies in multi-agent systems

Greg Hudas[1], Kyriakos G Vamvoudakis[2], Dariusz Mikulski[3] and Frank L Lewis[2]

Abstract

During mission execution in military applications, the TRADOC Pamphlet 525-66 Battle Command and Battle Space Awareness capabilities prescribe expectations that networked teams will perform in a reliable manner under changing mission requirements and changing team and individual objectives. In this paper we first present an overall view for dynamical decision-making in teams, both cooperative and competitive. Strategies for team decision problems, including optimal control, *N*-player games (H^∞ control, non-zero sum) and so on are normally solved offline by solving associated matrix equations such as the coupled Riccati equations or coupled Hamilton–Jacobi equations. However, using that approach, players cannot change their objectives online in real time without calling for a completely new offline solution for the new strategies. Therefore, in this paper we give a method for learning optimal team strategies online in real time as team dynamical play unfolds. In the linear quadratic regulator case, for instance, the method learns the coupled Riccati equations solution online without ever solving the coupled Riccati equations. This allows for truly dynamical team decisions where objective functions can change in real time and the system dynamics can be time-varying.

Keywords

multi-player games, Nash equilibrium, online gaming, optimal team behaviors, optimal team strategies.

1. Introduction

U.S. Army Training and Doctrine Command (TRADOC) Pamphlet 525-66 identifies Force Operating Capabilities required for the Army to fulfill its mission for a networked Warfighter concept. Two such capabilities are Battle Command and Battle-Space Awareness for which there are expectations that networked teams will perform in a reliable manner under changing mission requirements and changing team and individual objectives. These capabilities are necessary in the asymmetric battles waged against insurgencies, where enemy combatants quickly adapt to Army strategies and tactics.[1] This need is compounded by the fact that insurgents have increasingly become more difficult to detect due to their knowledge of the local terrain and their ability to mix with civilian populations. Over time, the needs of soldiers change in response to new insurgent strategies, quickly making existing technologies and systems obsolete. Since the Department of Defense currently does not have plans for fleet-wide upgrades for robots,[2,3] real-time adaptive team responses to insurgent threats are clearly key to mitigate the risk in uncertain and dynamic battle-spaces.

Battlefield or disaster area teams may be heterogeneous networks consisting of interacting humans, ground sensors, and unmanned airborne or ground vehicles (UAV, UGV). Such scenarios should provide real-time learning of optimal game strategies under changing mission requirements and team objectives. This requires adaptive algorithms for online learning of optimal solutions to multi-player games that facilitate keeping strategies updated as team and player

[1]U.S. Army RDECOM-TARDEC, Joint Center for Robotics (JCR), Warren, MI, USA
[2]Automation and Robotics Research Institute, University of Texas at Arlington, 7300 Jack Newell Boulevard South, Fort Worth, TX 76118, USA
[3]Oakland University, Oakland, MI, USA

Corresponding author:
Kyriakos G. Vamvoudakis, Automation and Robotics Research Institute, University of Texas at Arlington, 7300 Jack Newell Boulevard South, Fort Worth, TX 76118, USA.
Email: kyriakos@arri.uta.edu

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **SEP 2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Online adaptive learning for team strategies in multi-agent systems**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research, Development and Engineering Command (RDECOM),Tank Automotive Research, Development and Engineering Center (TARDEC),Joint Center for Robotics (JCR),Warren ,MI,48397**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

During mission execution in military applications, the TRADOC Pamphlet 525-66 Battle Command and Battle Space Awareness capabilities prescribe expectations that networked teams will perform in a reliable manner under changing mission requirements and changing team and individual objectives. In this paper we first present an overall view for dynamical decision-making in teams, both cooperative and competitive. Strategies for team decision problems, including optimal control, N-player games (H control, non-zero sum) and so on are normally solved offline by solving associated matrix equations such as the coupled Riccati equations or coupled Hamilton?Jacobi equations. However, using that approach, players cannot change their objectives online in real time without calling for a completely new offline solution for the new strategies. Therefore, in this paper we give a method for learning optimal team strategies online in real time as team dynamical play unfolds. In the linear quadratic regulator case, for instance, the method learns the coupled Riccati equations solution online without ever solving the coupled Riccati equations. This allows for truly dynamical team decisions where objective functions can change in real time and the system dynamics can be time-varying.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

objective functions change. Current methods of solving for optimal game strategies require offline solution of coupled matrix equations, which does not allow for straightforward updating of decision policies when objectives change. Required are deployable dynamic learning algorithms for keeping decision policies current to support mission tailoring, force responsiveness and agility, ability to change missions without exchanging forces, general adaptability to changing battlefield conditions, and defense against ballistic missile attack.[4]

This paper has two parts. First, it presents an overall view of team behaviors and dynamical decision-making in teams, both cooperative and competitive. In the second part, we show how to learn optimal game strategies online in real-time by observing data along the system trajectories as players interact with each other in cooperative or competitive play. In the first part we discuss cooperation, collaboration, altruistic vs. selfish behavior, antagonism, competition, incentives, minimum risk, cheating, and other concepts of multi-player team play. These concepts and others are rather easy to define clearly in terms of different objective/payoff functions, and/or different optimality criteria.

Strategies for team decision problems, including optimal control, *N*-player games (non- zero sum, zero sum) and so on are normally solved offline by solving the coupled Hamilton–Jacobi (HJ) equations for non-linear systems or coupled Riccati equations for linear systems. However, using that approach, players cannot change their objectives online in real time without calling for a completely new offline solution for the new strategies. Therefore, in the second part of this paper methods are given for solving different team decision problems online in real time by observing data along the system trajectories. This provides a truly dynamic framework for team decision-making, since players or teams can change their objectives or optimality criteria on the fly, and the new strategies for all players appropriate to the new situation can be re-computed in real time. This approach also allows for time-varying team dynamics.

The interplay between protagonists and opponents in team play is complex. Often, one's play is improved if one has to face an adversary. Our definitions of team behaviors in the first part of the paper were inspired by the summary of Chandler et al.[1] The discussions on dynamical games are based on Basar and Olsder.[5] A survey of *reinforcement learning* (RL) techniques for solving multi-player games is presented in an award winning paper by Busoniu et al.[6]

The approach given in the second part of the paper for online learning of optimal game solutions is based on concepts from RL.[7,8] Every living organism interacts with its environment and uses those interactions to improve its own actions in order to survive and increase. Charles Darwin showed that species modify their actions based on interactions with the environment over long time scales, leading to natural selection and survival of the fittest. RL refers to an actor or agent that interacts with its environment and modifies its actions, or control policies, based on stimuli received in response to its actions. RL implies goal directed behavior at least insofar as the agent has an understanding of reward versus lack of reward or punishment. Using a form of RL known as *policy iteration*,[8] we develop an algorithm for online learning of the solution to the *N*-player (zero-sum, non-zero-sum) game problem. In this algorithm, the optimal value of the game and the Nash equilibrium solution are learned in real-time as the players play together in a dynamical system scenario. The non-linear system case is presented. In the linear quadratic regulator special case, the algorithm learns the solution to coupled Riccati equations online, without ever actually solving the coupled Riccati equations.

Simulation examples show that the team learns the correct Nash equilibrium solution.

2. Different objectives and the behaviors of teams

The framework for team behaviors which we present in this paper can be applied for general non-linear multi-player dynamical systems, in continuous time or discrete time. We specialize to the linear time-invariant (LTI) continuous-time dynamical systems simply for ease of discussion. Therefore, consider the continuous-time LTI dynamical system given by

$$\dot{x} = A_x + B_1 u_1 + B_2 u_2 \tag{1}$$

with state $x(t) \in R^n$ and two control inputs or players $u_1(t) \in R^{m_1}$, $u_2(t) \in R^{m_2}$. The players may be cooperating or competing.

Define objective functions for players 1 and 2 respectively as

$$\begin{aligned} J_1(x(t), u_1, u_2) &= \tfrac{1}{2} \int_t^\infty L_1(x(\tau), u_1(\tau), u_2(\tau))\, d\tau \\ &= \tfrac{1}{2} \int_t^\infty (x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2)\, d\tau \end{aligned} \tag{2}$$

$$\begin{aligned} J_2(x(t), u_1, u_2) &= \tfrac{1}{2} \int_t^\infty L_1(x(\tau), u_1(\tau), u_2(\tau))\, d\tau \\ &= \tfrac{1}{2} \int_t^\infty (x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2)\, d\tau \end{aligned} \tag{3}$$

These are infinite horizon performances-to-go starting at time *t* in state *x*(*t*). They capture information equivalent to the payoff matrices of static games.[5]

In this paper one considers problems of *minimizing* the objective functions. That is, the objectives are considered as costs to be made small by proper selection of the players' strategies.

The integrands $L_i(x, u_1, u_2)$ are defined point wise at a time t in terms of weighting matrices Q and R and are known (loosely) as Lagrangians, or as utility functions. They are selected by the players or a higher-level authority depending on the performance requirements of the system. Interpreting the control inputs as feedback control strategies $u_1(x)$, $u_2(x)$, also called policies, that depend on the current system state $x(t)$, then $J_1(x(t), u_1(x), u_2(x))$, $J_2\,(x(t), u_1(x), u_2(x))$ represent the costs to players 1 and 2 respectively of motion along the system trajectories given the current strategies starting at time t in state $x(t)$. A variety of team decision problems and team behaviors can be defined through the choice of the objective functions.

2.1 Team coordination

We follow fairly closely the nice list of definitions[1] in discussing different sorts of team behaviors. *Coordination* is the closest and most cohesive form of cooperation in teams. There, all team members share a common objective function. That is,

$$\begin{aligned} J_1(x(t), u_1, u_2) &= J_2(x(t), u_1, u_2) \\ &= \tfrac{1}{2}\int_t^\infty (x^{\mathrm{T}}Qx + u_1^{\mathrm{T}}R_{11}u_1 + u_2^{\mathrm{T}}R_{12}u_2)\, d\tau \end{aligned} \tag{4}$$

All players have the same optimality criteria, namely to minimize the objective function. This is exactly the standard optimal control problem.[9] Military vehicle formations and convoys represent scenarios where all team members are obligated to participate and are bound to all assignments, tasks, or agreements.

The solution to this problem is given by solving the algebraic Riccati equation (ARE)

$$0 = A^{\mathrm{T}}P + PA + Q - PB_1R_{11}{}^{-1}B_1{}^{\mathrm{T}}P - PB_2R_{12}{}^{-1}B_2{}^{\mathrm{T}}P \tag{5}$$

and the feedback control policies are given by

$$u_1 = -K_1x \equiv -R_{11}{}^{-1}B_1{}^{\mathrm{T}}Px,\ \mathrm{u}_2 = -K_2x \equiv -R_{12}{}^{-1}B_2{}^{\mathrm{T}}Px \tag{6}$$

The Riccati equation solution must be performed offline a priori, and it defines the feedback policies for all time. Unfortunately, this rules out the possibility of truly dynamic team behavior since the utility function weighting matrices cannot be changed on the fly in real time. We show how to fix this in Section 4 through online learning of the optimal strategies.

2.2 Team cooperation and collaboration

Cooperation is a looser form of team cohesiveness whereby each player can have its own objective function, as given in (2) and (3), in addition to team objective functions. If each player seeks to minimize their own cost function, the solution to this optimal control problem is given[5] in terms of the two coupled AREs

$$\begin{aligned} 0 = A_c{}^{\mathrm{T}}P_1 + P_1A_c + Q_1 + P_1B_1R_{11}{}^{-1}B_1{}^{\mathrm{T}}P_1 + \\ P_1B_2R_{22}{}^{-1}\mathrm{R}_{12}R_{22}{}^{-1}B_2{}^{\mathrm{T}}P_1 \end{aligned} \tag{7}$$

$$\begin{aligned} 0 = A_c{}^{\mathrm{T}}P_2 + P_2A_c + Q_2 + P_2B_1R_{11}{}^{-1}R_{21}R_{11}{}^{-1}B_1{}^{\mathrm{T}}P_2 + \\ P_2B_2R_{22}{}^{-1}B_2{}^{\mathrm{T}}P_2 \end{aligned} \tag{8}$$

where the closed-loop system matrix is

$$A_c = A - B_1K_1 - B_2K_2 \tag{9}$$

and the optimal feedback policies are given by

$$u_1 = -K_1x \equiv -R_{11}{}^{-1}B_1{}^{\mathrm{T}}P_1x,\ \mathrm{u}_2 = -K_2x \equiv -R_{22}{}^{-1}B_2{}^{\mathrm{T}}P_2x \tag{10}$$

Collaboration is a still looser form of team behavior whereby each player seeks to optimize their objective function without compromising team task completion. The spectrum between coordination, cooperation, and collaboration is a continuum that depends on the closeness of the objective functions of individual players. For example, the behavior of an Army medic and a wounded Soldier can be modeled by cooperative solutions, since both have different private objective functions, but share a team objective to move to a safe zone. A Predator drone, on the other hand, is more likely to exhibit collaborative behaviors for intelligence, surveillance, and reconnaissance (ISR), as long as it has enough fuel to fly.

To reflect the greater cohesiveness in cooperation than in collaboration, Chandler et al.[1] suggest defining a team objective function $J_{team}(x(t), u_1, u_2)$ and then setting

$$\begin{aligned} J_1(x(t), u_1, u_2) = (1{-}w_1)J_{team}(x(t), u_1, u_2) + \\ w_1J_{1p}\,(x(t), u_1, u_2) \end{aligned} \tag{11}$$

$$\begin{aligned} J_2(x(t), u_1, u_2) = (1{-}w_2)J_{team}(x(t), u_1, u_2) + \\ w_2J_{2p}\,(x(t), u_1, u_2) \end{aligned} \tag{12}$$

where J_{1p}, J_{2p} are private objectives for each player and w_1, w_2 are weightings selected to put more or less emphasis on team objectives as compared with private objectives.

2.3 Competition and conflict: zero-sum games

The resources available for survival or operation are often limited. Different players or different teams may compete against each other for the same limited resources, such as bandwidth on communication networks or natural resources such as land. The extreme case of *competitive behavior* is when

$$-J_2(x(t),u_1,u_2) = J_1(x(t),u_1,u_2) = \tfrac{1}{2}\int_t^{\infty}(x^{\mathrm{T}}Qx + u_1^{\mathrm{T}}R_{11}u_1 - u_2^{\mathrm{T}}R_{12}u_2)\,d\tau \quad (13)$$

That is, when one player wins, the other loses by the same amount. This is the standard *two-player zero-sum game*.[5,10] If both players seek to minimize their respective costs, the optimal solution is given by the solution to the game (or generalized) ARE

$$0 = A^{\mathrm{T}}P + PA + Q - PB_1R_{11}{}^{-1}B_1{}^{\mathrm{T}}P + PB_2R_{12}{}^{-1}B_2{}^{\mathrm{T}}P \quad (14)$$

with the optimal feedback strategies given by

$$u_1 = -K_1x \equiv -R_{11}{}^{-1}B_1{}^{\mathrm{T}}Px,\ \mathrm{u}_2 = K_2x \equiv R_{12}{}^{-1}B_2{}^{\mathrm{T}}Px \quad (15)$$

The solution to the two-player zero-sum game also provides the solution to the bounded L_2-gain problem, wherein the control input $u_1(t)$ seeks to guarantee bounds on the output in the face of a system disturbance given by $u_2(t)$. In this context one selects $R_{12} = \gamma^2$ for a fixed scalar $\gamma > 0$ and the guaranteed bound is given in terms of L_2 function norms as

$$\| z \| \le \gamma \| u_2 \|,\ z = \begin{bmatrix} \sqrt{Q}\,x \\ \sqrt{R_{11}}\,u_1 \end{bmatrix} \quad (16)$$

with $z(t)$ the performance output. Under reachability and observability conditions this solution exists and is unique for large enough $\gamma > 0$.

2.4 Decomposition of objective functions into team goals plus conflict of interest goals

The objective functions of each player can be written as a *team average* term plus a *conflict of interest* term.

For the case of two players one has

$$J_1 = \tfrac{1}{2}(J_1 + J_2) + \tfrac{1}{2}(J_1 - J_2) \equiv J_{\mathrm{team}} + J_1^{\mathrm{coi}} \quad (17)$$

$$J_2 = \tfrac{1}{2}(J_1 + J_2) + \tfrac{1}{2}(J_2 - J_1) \equiv J_{\mathrm{team}} + J_2^{\mathrm{coi}} \quad (18)$$

For three players

$$J_1 = \tfrac{1}{3}(J_1 + J_2 + J_3) + \tfrac{1}{3}(J_1 - J_2) + \tfrac{1}{3}(J_1 - J_3) \equiv J_{\mathrm{team}} + J_1^{\mathrm{coi}} \quad (19)$$

$$J_2 = \tfrac{1}{3}(J_1 + J_2 + J_3) + \tfrac{1}{3}(J_2 - J_1) + \tfrac{1}{3}(J_2 - J_3) \equiv J_{\mathrm{team}} + J_2^{\mathrm{coi}} \quad (20)$$

$$J_3 = \tfrac{1}{3}(J_1 + J_2 + J_3) + \tfrac{1}{3}(J_3 - J_1) + \tfrac{1}{3}(J_3 - J_2) \equiv J_{\mathrm{team}} + J_3^{\mathrm{coi}} \quad (21)$$

For N players one may write

$$J_i = \tfrac{1}{N}\sum_{j=1}^{N} J_j + \tfrac{1}{N}\sum_{j=1}^{N}(J_i - J_j) \equiv J_{\mathrm{team}} + J_i^{\mathrm{coi}},\ i=1, N \quad (22)$$

For *N-player zero-sum games*, the first term is zero, i.e. the players have no goals in common. The case of zero-sum multi-player games in a competition mode is discussed by Busoniu et al.[6]

3. Different optimality criteria and the behaviors of teams

The behaviors of teams and individual players change depending on the selection of the objective functions. Given the multi-player games just described, one can have further differing team behaviors depending on the prescribed optimality criteria, and also on the definition of equilibrium point.

3.1 Nash Equilibria and Myopic Self-Improvement

Let each player seek to minimize their own objective function. The *Nash equilibrium policy*[5] (u_1^*, u_2^*) for a two-player game is defined by the conditions

$$J_1(x,u_1^*,u_2^*) \le J_1(x,u_1,u_2^*)$$

$$J_2(x,u_1^*,u_2^*) \le J_2(x,u_1^*,u_2) \quad (23)$$

This means that if either player changes their own strategy while the other does not, they do worse in terms of having an increased cost. Nash equilibria are stable in the sense that a single player cannot improve their performance by unilateral actions. Each player considers only their own *selfish cost*. Under certain standard conditions,[5] Nash equilibria exist and are unique.

In the context of two-player zero-sum games, the optimality criterion can be expressed as

$$J_1(x(t),u_1^*,u_2^*) = \tfrac{1}{2}\min_{u_1}\max_{u_2}\int_t^{\infty}(x^{\mathrm{T}}Qx + u_1^{\mathrm{T}}R_{11}u_1 - u_2^{\mathrm{T}}R_{12}u_2)\,d\tau \quad (24)$$

whereby player 1 seeks to minimize the objective function while player 2 seeks to maximize it. Then the Nash condition can be written as

$$J_1(x,u_1^*,u_2) \le J_1(x,u_1^*,u_2^*) \le J_1(x,u_1,u_2^*) \quad (25)$$

3.2 Pareto equilibria, agreements, and cheating

A different sort of optimality criterion is defined by the *Pareto equilibrium*, which for two players satisfies the conditions

$$\text{if } J_1(x,u_1,u_2^*) < J_1(x,u_1^*,u_2^*), \text{then } J_2(x,u_1^*,u_2^*) < J_2(x,u_1,u_2^*)$$

$$\text{if } J_2(x,u_1^*,u_2) < J_2(x,u_1^*,u_2^*),$$

$$\text{then } J_1(x,u_1^*,u_2^*) < J_1(x,u_1^*,u_2) \quad (26)$$

This means that if either player adopts a strategy other than the equilibrium, either they will incur increased costs or other players will. This is an *altruistic sense* of equilibrium wherein all *players seek to help team members* improve their performance. This particular optimality criterion aligns well to the Army Core Values of loyalty and self-service.

Pareto equilibria are not necessarily unique. To obtain a unique equilibrium, additional *side agreements* are needed between the players. Moreover, Pareto equilibria require *cooperation* between players and an agreement that none will act so as to harm another. Pareto performance is subject to *cheating*, defined as a situation where at least one player does not follow the agreed-upon rules. By cheating, a single player may be able to improve their performance at the expense of their team mates.

3.3 Antagonistic behavior

Another sort of equilibrium is defined by the conditions

$$J_2(x,u_1^*,u_2^*) \geq J_2(x,u_1,u_2^*)$$
$$J_1(x,u_1^*,u_2^*) \geq J_1(x,u_1^*,u_2) \tag{27}$$

This means that if player 1 diverges from the equilibrium policy, then player 2 will have an improved payoff in terms of decreased costs, and vice versa. That is, each player is interested in harming the other as much as possible. This is a definition of *antagonistic behavior*. The uniqueness of such equilibria needs to be established.

3.4 Leader–follower games and team incentives

Consider the case of two players and define the equilibrium solution as

$$J_2(x,u_1,u_2^*) \leq J_2(x,u_1,u_2) \text{ for a fixed policy } u_1(x)$$
$$J_1(x,u_1^*,u_2^*) \leq J_1(x,u_1,u_2^*) \tag{28}$$

This is a *hierarchical decision problem* in which player 1 acts as the *leader* and player 2 as the *follower*. The objective of lead player 1 is to determine an *incentive*, through selection of their policy $u_1(x)$, so that the follower will always play so as to minimize the leader's cost while seeking to minimize their own. This is known as a *Stackelberg game*.

In the case of three or more players, one can have several definitions of equilibrium point.[5] Consider the definition for three players given by

$$J_2(x,u_1,u_2^*,u_3^*) \leq J_2(x,u_1,u_2,u_3^*) \text{ for a fixed policy } u_1(x) \tag{29}$$

$$J_3(x,u_1,u_2^*,u_3^*) \leq J_3(x,u_1,u_2^*,u_3) \text{ for a fixed policy } u_1(x) \tag{30}$$

$$J_1(x,u_1^*,u_2^*,u_3^*) \leq J_3(x,u_1,u_2^*,u_3^*) \tag{31}$$

In this situation, players 2 and 3 are followers who adopt a Nash equilibrium with regards to each other in the *followers subgame* for each policy of the leader. The objective of lead player 1 is to determine an incentive so that the followers will always act to minimize their cost while seeking to minimize its own.

Stackelberg strategies have been explored in the context of international terrorism[11] and show their obvious implications within existing government/military leadership hierarchies. John Keegan, who wrote a history of men at war in *The Face of Battle*, talks about how 'the personal bond between leader and follower lies at the root of all explanations of what does and does not happen in battle'. This quote eloquently describes how the incentive for soldiers to follow orders from their superiors can seriously affect both soldier morale and wartime outcomes.

4. Online learning of optimal team strategies

It has been shown[6] in cooperative games that the agents use the same objective function and they use greedy policies to maximize their common return. Furthermore a variety of team and individual player strategies can be defined by suitable selection of payoff objective functions and suitable definitions of optimality. Normal approaches to solving for optimal strategies for team decision problems involve offline solution, such as the coupled Riccati equations. However, using that approach, players cannot change their objectives online in real time without calling for a completely new offline solution for the new strategies. In this section we show how to compute optimal team strategies online in real time by learning based on observed data along the system trajectories. This provides a truly dynamic framework for team decision making, since players or teams can change their objectives or optimality criteria on the fly, and the new strategies for all players appropriate to the new situation are then re-computed in real time. This online gaming approach also allows for time-varying team dynamics.

This learning approach is based on RL techniques.[7,8] A survey on multi-agent RL is presented by Busoniu et al.[6] It is a general method for solving optimal decision problems for *general non-linear* dynamical systems, and will be illustrated for the non-linear two-player game solution (zero-sum or non-zero sum).

4.1 N-player non-linear games

Consider the *N*-player non-linear time-invariant differential game on an infinite time horizon

$$\dot{x} = f(x) + \sum_{j=1}^{N} g_j(x)u_j \tag{32}$$

where state $x(t) \in R^n$, controls $u_j(t) \in R^{m_j}$. Assume that $f(x)$ is continuously differentiable and $f(0) = 0$ so that $x = 0$ is an equilibrium point of the system. The cost functionals associated with each player are[2]

$$J_i(x(0), u_1, u_2, \ldots u_N) = \int_0^\infty (Q_i(x) + \sum_{j=1}^N u_i^{\mathrm{T}} R_{ij} u_i)\, dt \equiv \int_0^\infty r_i(x(t), u_1, u_2, \ldots u_N)\, dt; \quad i \in N \tag{33}$$

where function $Q_\mathrm{i}(x) \geq 0$ is generally non-linear, and $R_{ii} > 0$, $R_{ij} > 0$ are symmetric matrices.

Given admissible feedback policies/strategies $u_i(t) = \mu_i(x)$ the value is

$$V_i(x(0), \mu_1, \mu_2, \ldots \mu_N) = \int_t^\infty (Q_i(x) + \sum_{j=1}^N \mu_i^{\mathrm{T}} R_{ij} \mu_i)\, d\tau \equiv \int_t^\infty r_i(x(t), \mu_1, \mu_2, \ldots \mu_N)\, d\tau; \; i \in N \tag{34}$$

Define the N-player game $V_i^*(x(t), \mu_1, \mu_2, \ldots \mu_N) =$

$$\min_{\mu_i} \int_t^\infty (Q_i(x) + \sum_{j=1}^N \mu_i^{\mathrm{T}} R_{ij} \mu_i)\, d\tau; \quad i \in N \tag{35}$$

By assuming that all of the players have the same hierarchical level, we focus on the so-called Nash equilibrium that is given by the following definition.

Definition 1. (Nash equilibrium strategies, Başar and Olsder[5]) An *N-tuple* of strategies $\{\mu_1^*, \mu_2^*, ..., \mu_\mathrm{N}^*\}$ $\mu_i^* \in \Omega_i$, $i \in N$ with $\mu_i^* \in \Omega_i$, $i \in N$ is said to constitute a Nash equilibrium solution for an *N-player* finite game in extensive form, if the following N inequalities are satisfied for all $\mu_i^* \in \Omega_i$, $i \in N$:

$$\left.\begin{array}{c} J_1^* \triangleq J_1(\mu_1^*, \mu_2^*, ..., \mu_\mathrm{N}^*) \leq J_1(\mu_1, \mu_2^*, ..., \mu_\mathrm{N}^*) \\ J_2^* \triangleq J_2(\mu_1^*, \mu_2^*, ..., \mu_\mathrm{N}^*) \leq J_2(\mu_1^*, \mu_2, ..., \mu_\mathrm{N}^*) \\ \ldots \\ \ldots \\ J_N^* \triangleq J_N(\mu_1^*, \mu_2^*, ..., \mu_\mathrm{N}^*) \leq J_N(\mu_1, \mu_2^*, ..., \mu_\mathrm{N}) \end{array}\right\} \tag{36}$$

The *N-tuple* of quantities $\{J_1^*, J_2^*, ..., J_\mathrm{N}^*\}$ is known as a Nash equilibrium outcome of the N-player game.

Differential equivalents to each value function are given by the following non-linear Lyapunov equations

$$0 = r(x, u_1, \ldots, u_N) + (\nabla V_i)^{\mathrm{T}} (f(x) + \sum_{j=1}^N g_j(x) u_j), \quad V_i(0) = 0, \quad i \in N \tag{37}$$

where $\nabla V_i = \partial V_i / \partial x \in R^{n_i}$ is the gradient vector (e.g. transposed gradient). Then, suitable non-negative-definite solutions to (37) are the values evaluated using the infinite integral (35) along the system trajectories. Define the Hamiltonian functions

$$H_i(x, \nabla V_i, u_1, \ldots, u_N) = r(x, u_1, \ldots, u_N) + (\nabla V_i)^{\mathrm{T}} (f(x) + \sum_{j=1}^N g_j(x) u_j) \quad i \in N \tag{38}$$

According to the stationarity conditions, associated feedback control policies are given by

$$\frac{\partial H_i}{\partial u_i} = 0 \Rightarrow \mu_i(x) = -\tfrac{1}{2} R_{ii}^{-1} g_i^{\mathrm{T}}(x) \nabla V_i, \quad i \in N \tag{39}$$

Substituting (39) into (37) one obtains the N coupled HJ equations

$$0 = (\nabla V_i)^{\mathrm{T}} \left(f(x) - \tfrac{1}{2} \sum_{j=1}^N g_j(x) R_{jj}^{-1} g_j^{\mathrm{T}}(x) \nabla V_j \right) + Q_i(x) + \tfrac{1}{4} \sum_{j=1}^N \nabla V_j{}^{\mathrm{T}} g_j(x) R_{jj}^{-\mathrm{T}} R_{ij} R_{jj}^{-1} g_j^{\mathrm{T}}(x) \nabla V_j, \quad V_i(0) = 0 \tag{40}$$

These coupled HJ equations are in 'closed-loop' form. The equivalent 'open-loop' form is

$$0 = \nabla V_i^{\mathrm{T}} f(x) + Q_i(x) - \tfrac{1}{2} \nabla V_i^{\mathrm{T}} \sum_{j=1}^N g_j(x) R_{jj}^{-1} g_j^{\mathrm{T}}(x) \nabla V_j + \tfrac{1}{4} \sum_{j=1}^N \nabla V_j^{\mathrm{T}} g_j(x) R_{jj}^{-\mathrm{T}} R_{ij} R_{jj}^{-1} g_j^{\mathrm{T}}(x) \nabla V_j, \quad V_i(0) = 0 \tag{41}$$

These equations are difficult to solve. An iterative offline solution technique is given by the Policy Iteration algorithm in the next section and it is the key to motivate the control structure for an online adaptive N-player game solution algorithm. Then it is proven that 'optimal adaptive' control algorithm converges online to the solution of coupled HJs (41), while guaranteeing closed-loop stability.

4.2 Solution of the N-player game using reinforcement learning

The optimal strategies of the N-player game are given in terms of the coupled HJ equations (41). Unfortunately, the coupled HJ equations (41) are usually intractable to solve directly. In fact, the coupled HJ equations may not have exact analytic solutions. One can solve the coupled HJ equations iteratively to obtain a suitable local smooth solution by using one of several algorithms built on techniques from RL.[8] One method of RL is known as *policy iteration*. The following policy iteration algorithm solves the coupled HJ equations by iterative solution of a far simpler equation, namely the non-linear Lyapunov-like equations (37).

4.2.1 Policy iteration for N-player games

Start with stabilizing initial policies $\mu^0_{\,1}(x)$,..., $\mu^0_{\,N}(x)$. Given the N-tuple of policies , solve for the N-tuple of costs $V^k{}_1(x(t)), V^k{}_2(x(t)) \ldots V^k{}_N(x(t))$ using

$$0 = r(x,\mu^k{}_1,\ldots,\mu^k{}_N) + (\nabla V_i^k)^{\mathrm{T}} \left(f(x) + \sum_{j=1}^{N} g_j(x)\mu_j^i \right), \quad V^k{}_i(0) = 0 \quad i \in N \tag{42}$$

Update the N-tuple of control policies using

$$\mu_i^{k+1} = \underset{u_i \in \Psi(\Omega)}{\operatorname{argmin}} [H_i(x, \nabla V_i, u_1, \ldots, u_N)] \quad i \in N \tag{43}$$

which explicitly is

$$\mu_i^{k+1}(x) = -\tfrac{1}{2} R_{ii}^{-1} g_i^{\mathrm{T}}(x) \nabla V_i^k \quad i \in N \tag{44}$$

A linear version of the previous algorithm is presented by Gajic and Li[12] and Abou-Kandil et al.[13]

The PI algorithm will be used as the basis for online learning solution techniques for optimal game strategies in the next section.

4.3 Online gaming solution of the two-player game

The PI Algorithm is a *sequential* algorithm that solves the coupled HJ equations (42) and finds the optimal strategies (44) for the game. In this section, we develop an *online algorithm for learning the solution to the two-player differential game in real time.* In this algorithm, the two players *learn simultaneously* as they play together in a dynamical game. This is in effect an adaptive control algorithm of novel form that converges to the optimal game solution. The online gaming algorithm is motivated by the PI algorithm. Consider the non-linear dynamical system given by

$$\dot{x} = f(x) + g(x)u + k(x)d \tag{45}$$

with state $x(t) \in R^n$, first control $u(t) \in R^m$, and second control $d(t) \in R^q$. Assume that $f(x)$ is continuously differentiable and $f(0) = 0$ so that $x = 0$ is an equilibrium point of the system.

The online gaming algorithm is an adaptive learning controller that is based on *value function approximation* (VFA).[14-16] Motivated by Equations (42) and (43) in the PI algorithm, it uses four approximator structures, which can be considered as neural networks (NNs). Two NNs learn the current values of the game, i.e. the solution of the non-linear Lyapunov equations (42) for the current control policies. The other two NNs learn the two control policies.

That is, one has the estimates of the values, and control policies respectively expressed as

$$\hat{V}_1(x) = \hat{W}_1^{\mathrm{T}} \phi_1(x),\ \hat{V}_2(x) = \hat{W}_2^{\mathrm{T}} \phi_2(x) \tag{46}$$

$$u_3(x) = -\tfrac{1}{2} R_{11}{}^{-1} g^{\mathrm{T}}(x) \nabla \phi_1^{\mathrm{T}} \hat{W}_3 \tag{47}$$

$$d_4(x) = -\tfrac{1}{2} R_{22}{}^{-1} k^{\mathrm{T}}(x) \nabla \phi_2^{\mathrm{T}} \hat{W}_4 \tag{48}$$

Here, the weights of the four NNs are $\hat{W}_1, \hat{W}_2, \hat{W}_3, \hat{W}_4$. Exactly as in adaptive control, these are matrices of unknown parameters which must be estimated or tuned by online learning methods. The NN activation functions are $\phi_1(x)$, $\phi_2(x)$ and $\nabla\phi_1 = \partial\phi_1/\partial x$, $\nabla\phi_2 = \partial\phi_2/\partial x$ are the Jacobian matrices.

This scheme has the so-called actor–critic structure,[8,15,16] whereby the critic NNs (46) seek to learn the values of the current policies, e.g. the solution to the non-linear Lyapunov equations (42). The actor NNs (47) and (48), on the other hand, seek to learn the optimal policies for both players. The main theorem is now given. It provides the tuning laws for the critic, and control NNs that guarantee convergence of the online gaming algorithm in real-time to the Nash equilibrium solution, while guaranteeing closed-loop stability.

Theorem 1. (Online games) Let the dynamics for the two-player game be given by (45), and consider the game formulation as analyzed in this section. Let the critic NNs be given by (46), the first control input be given by actor (first player) NN (47) and the second control input be given by actor (second player) NN (48). Let tuning for the first critic NN be provided by

$$\dot{\hat{W}}_1 = -a_1 \frac{\sigma_3}{(\sigma_3^{\mathrm{T}}\sigma_3 + 1)^2} [\sigma_3{}^{\mathrm{T}} \hat{W}_1 + Q_1(x) + u_3{}^{\mathrm{T}} R_{11} u_3 + d_4{}^{\mathrm{T}} R_{12} d_4] \tag{49}$$

and the second critic NN be provided by

$$\dot{\hat{W}}_2 = -a_2 \frac{\sigma_4}{(\sigma_4^{\mathrm{T}}\sigma_4 + 1)^2} [\sigma_4{}^{\mathrm{T}} \hat{W}_2 + Q_2(x) + u_3{}^{\mathrm{T}} R_{21} u_3 + d_4{}^{\mathrm{T}} R_{22} d_4] \tag{50}$$

where $\sigma_3 = \nabla\phi_1(f + gu_3 + kd_4)$ and $\sigma_4 = \nabla\phi_2(f + gu_3 + kd_4)$. Let the first actor NN (first player) be tuned as

$$\dot{\hat{W}}_3 = -\alpha_3 \Big\{ (F_2 \hat{W}_3 - F_1 \bar{\sigma}_3^{\mathrm{T}} \hat{W}_1) - \frac{1}{4} \nabla\phi_1 g(x) R_{11}{}^{-\mathrm{T}} R_{21} R_{11}{}^{-1} g^{\mathrm{T}}(x) \nabla\phi_1{}^{\mathrm{T}} \hat{W}_3 m_2{}^{\mathrm{T}} \hat{W}_2 - \frac{1}{4} \bar{D}_1(x) \hat{W}_3 m_1^T \hat{W} \Big\} \tag{51}$$

and the second actor (second player) NN be tuned as

$$\dot{\hat{W}}_4 = -\alpha_4 \Big\{ \big(F_4 \hat{W}_4 - F_3 \bar{\sigma}_4^T \hat{W}_2 \big) - \frac{1}{4} \nabla\phi_2 k(x) R_{22}{}^{-T} R_{12} R_{22}{}^{-1} k^T(x) \nabla\phi_2{}^T \hat{W}_4 m_1^T \hat{W} - \frac{1}{4} \bar{D}_2(x) \hat{W}_4 m_2^T \hat{W}_2 \Big\} \tag{52}$$

where $\bar{D}_1(x) \equiv \nabla\phi_1(x) g(x) R_{11}{}^{-1} g^{\mathrm{T}}(x) \nabla\phi_1^{\mathrm{T}}(x)$,
$\bar{D}_2(x) \equiv \nabla\phi_2(x) k R_{22}{}^{-1} k^{\mathrm{T}} \nabla\phi_2^{\mathrm{T}}(x)$,

$m_1 \equiv \dfrac{\sigma_3}{(\sigma_3^T \sigma_3 + 1)^2}$, $m_2 \equiv \dfrac{\sigma_3}{(\sigma_4^T \sigma_4 + 1)^2}$ and $F_1 > 0$, $F_2 < 0$, $F_3 > 0$, $F_4 > 0$ are tuning parameters. Also assume $Q_1(x) > 0$ and $Q_2(x) > 0$. Then there exists an N_0 such that, for the number of NN hidden layer units $N > N_0$ the closed-loop system state, the critic NN errors $\tilde{W}_1, \tilde{W}_2$, and the actor NN errors $\tilde{W}_3, \tilde{W}_4$ are bounded. Moreover, $\hat{V}_1(x)$ and $\hat{V}_2(x)$ converge to the solution to the coupled HJ equations.

It is important to note that this algorithm learns the solution to the coupled HJ equations (42) and the optimal policies (44) without ever in fact solving either the coupled HJ equations or the Lyapunov equations. In the LQR case, it learns the solution to the coupled Riccati equations online.

5. Simulations

5.1 Two player non-linear system

In 1954, Colonel O.G. Haywood asserted[17] the little known fact that von Neumann's minorant solution to two-player zero-sum games is identical to the U.S. Military decision doctrine known as the 'Estimate of the Situation'. Today, two-player zero-sum games are viewed as essential tools for military commanders to determine the optimal solution in an uncertain wartime situation.[18] Two-player non-zero-sum games are equally important when developing strategies for civilian–military cooperation in peacetime operations.[19] These examples illustrate how two-player system models are relevant to military operations, even when their usage involves teams or groups rather than just individual agents.

Consider the following affine in control, two-player (u and d) non-linear system, with a quadratic cost constructed as[20,21]

$$\dot{x} = f(x) + g(x)u + k(x)d, x \in R^2$$

where

$$f(x) = \begin{bmatrix} x_2 \\ -x_2 - 0.5x_1 + 0.25x_2(\cos(2x_1)+2)^2 + 0.25x_2(\sin(4x_1^2)+2)^2 \end{bmatrix}$$
$$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix}, \quad k(x) = \begin{bmatrix} 0 \\ \sin(4x_1^2)+2 \end{bmatrix}.$$

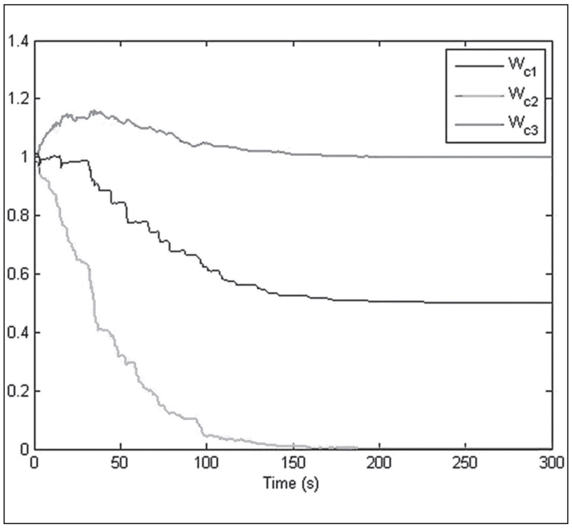

Figure 1. Convergence of the critic parameters for the first player

Select $Q_1 = 2Q_2$, $R_{11} = 2R_{22}$ and $R_{12} = 2R_{21}$, where Q_1, R_{22} and R_{21} are identity matrices.

The optimal value function for the first critic (player 1) is $V_1^*(x) = \frac{1}{2}x_1^2 + x_2^2$ and for the second critic (player 2) is $V_2^*(x) = \frac{1}{4}x_1^2 + \frac{1}{2}x_2^2$.

The optimal control signal for the first player is $u^*(x) = -(\cos(2x_1) + 2)x_2$ and the optimal control signal for the second player is $d^*(x) = -(\sin(4x_1^2) + 2)x_2$.

One selects the NN vector activation function for the critics as $\varphi_1(x) = \varphi_2(x) \equiv [x_1^2 \quad x_1x_2 \quad x_2^2]$. Figure 1 shows

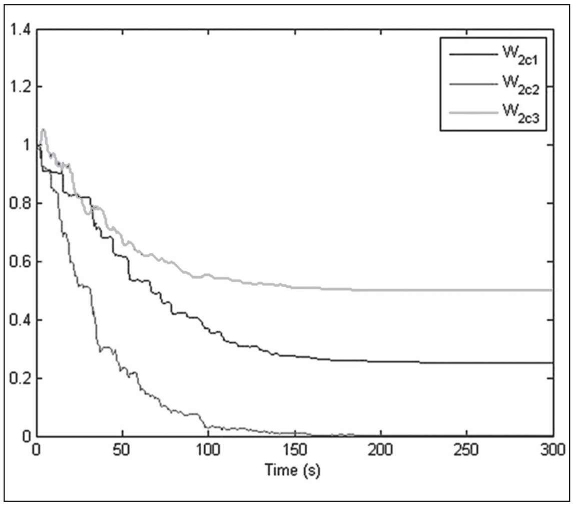

Figure 2. Convergence of the critic parameters for the second player

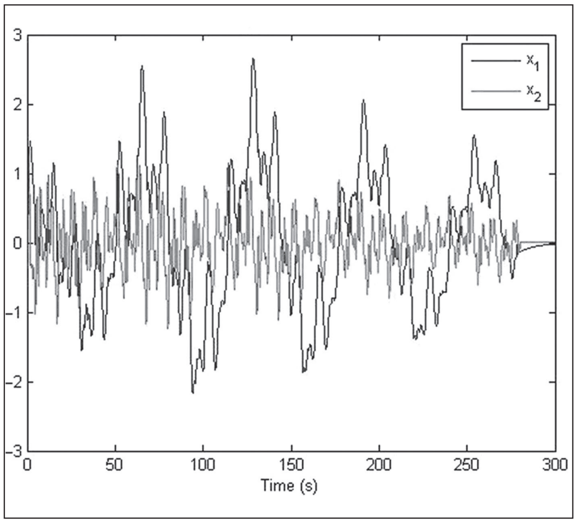

Figure 3. Evolution of the system states

the critic parameters for the first player, denoted by $\hat{W}_1 = [W_{c1}\ \ W_{c2}\ \ W_{c3}]^{\mathrm{T}}$ by using the proposed game algorithm. After convergence at about 150 s one has $\hat{W}_1(t_f) = [0.5015\ \ 0.0007\ \ 1.0001]^{\mathrm{T}}$.

Figure 2 shows the critic parameters, denoted by $\hat{W}_2 = [W_{2c1}\ \ W_{2c2}\ \ W_{2c3}]^{\mathrm{T}}$ by using the proposed game algorithm. After convergence at about 150 s one has $\hat{W}_2(t_f) = [0.2514\ \ 0.0006\ \ 0.5001]^{\mathrm{T}}$.

The actor parameters for the first player after 150 s converge to the values of $\hat{W}_3(t_f) = [0.5015\ \ 0.0007\ \ 1.0001]^{\mathrm{T}}$ and the actor parameters for the second player after 150 s converge to the values of $\hat{W}_4(t_f) = [0.2514\ \ 0.0006\ \ 0.5001]^{\mathrm{T}}$ Therefore the actor NN for the first player

$$\hat{u}(x) = -\tfrac{1}{2}R_{11}{}^{-1}\begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0.5015 \\ 0.0007 \\ 1.0001 \end{bmatrix}$$

also converged to the optimal control, and the actor NN for the second player

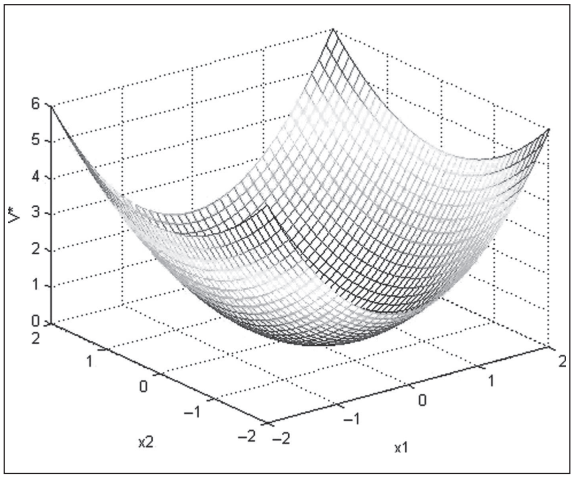

Figure 4. Optimal value function for player 1

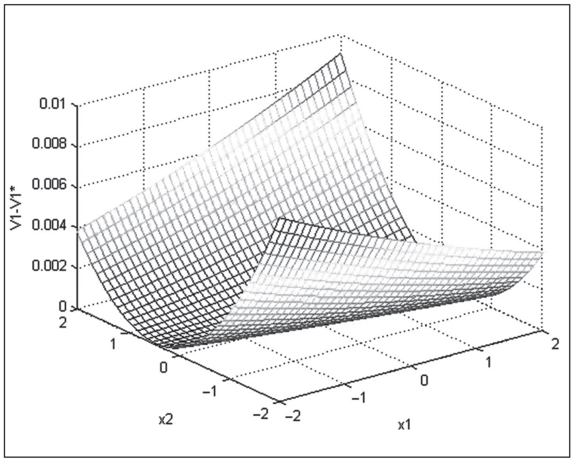

Figure 5. 3D plot of the approximation error for the value function of player 1.

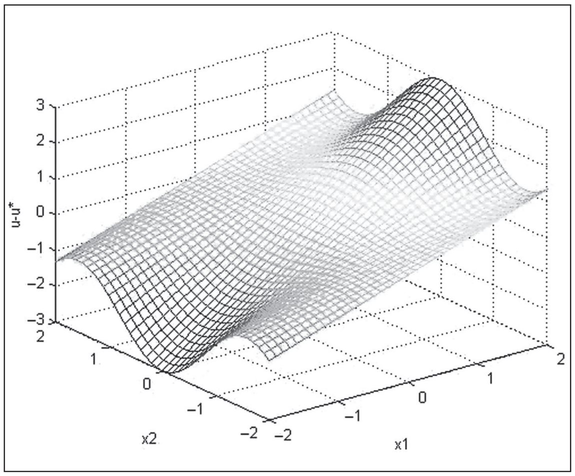

Figure 6. 3D plot of the approximation error for the control of player 1.

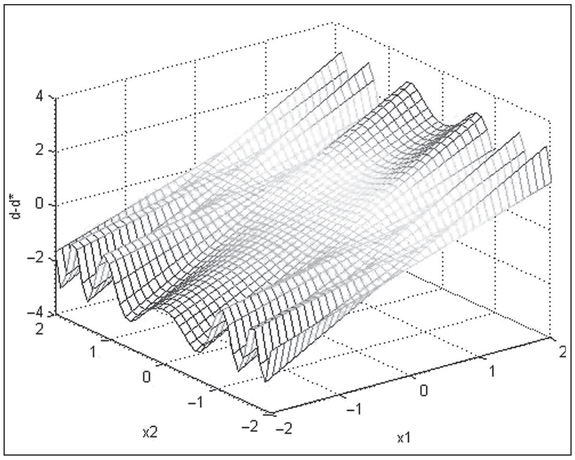

Figure 7. 3D plot of the approximation error for the control of player 2.

$$\hat{d}(x) = -\tfrac{1}{2} R_{22}^{-1} \begin{bmatrix} 0 \\ \sin(4x_1^2) + 2 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix}^T \begin{bmatrix} 0.2514 \\ 0.0006 \\ 0.5001 \end{bmatrix}$$

also converged to the optimal one.

The evolution of the system states is presented in Figure 3, where one can see how the PE influences the states.

Figure 4 shows the optimal value function for player 1 (similarly for player 2).

Figure 5 shows the 3D plot of the difference between the approximated value function for player 1 and the optimal one. Player 2 has a similar error. These errors are close to zero.

Good approximations of the actual value functions are being evolved. Figure 6 shows the 3D plot of the difference between the approximated control for the first player, by using the online algorithm, and the optimal one. This error is close to zero. Same for the second player in Figure 7.

Funding

This work was supported by ARO grant W91NF-05-1-0314, the National Automotive Center NAC, NSF grant ECCS-0801330, and AFOSR grant FA9550-09-1-0278.

References

1. Chandler P and Pachter M. Challenges. In Shima T and Rasmussen S (eds), *UAV Cooperative Decision and Control*. Philadelphia, PA: SIAM, 2009.
2. Erwin SI. Defense Dept. forecasts great use of robots in ground combat. *National Defense*. 2009; April: 240–25.
3. Singer PW. “Advanced” warfare: how we might fight with robots. In *Wired For War*. New York: The Penguin Press.
4. Palmore J and Melese F. A game theory view of preventive defense against ballistic defense attack. *Defense and Security Analysis* 2001; 17: 211–215.
5. Başar T and Olsder GJ. *Dynamic Noncooperative Game Theory*, 2nd ed (*SIAM’s Classic Series in Applied Mathematics*, Vol. 23). Philadelphia, PA: SIAM, 1999.
6. Busoniu L, Babuska R and De Schutter B. A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews* 2008; 38: 156–172.
7. Lewis FL and Vrabie D. Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control. *IEEE Circuits and Systems Magazine* 2009; 9(3): 32–50.
8. Sutton RS and Barto AG. *Reinforcement Learning – An Introduction*. Cambridge, MA: MIT Press, 1998.
9. Lewis FL and Syrmos VL. *Optimal Control*. New York: John Wiley & Sons, Inc., 1995.
10. Başar T and Bernard P. *H_∞ Optimal Control and Related Minimax Design Problems*. Boston, MA: Birkhäuser, 1995.
11. Behrens DA, Caulkins JP, Feichtinger G and Tragler G. Incentive Stackelberg strategies for a dynamic game on terrorism. *Adv Dynam Game Theory* 2007; 9: 459–486.
12. Gajic Z and Li T-Y. Simulation results for two new algorithms for solving coupled algebraic Riccati equations. In *Third International Symposium on Differential Games*, Sophia, Antipolis, France, 1988.
13. Abou-Kandil H, Freiling G, Ionescu V and Jank G. *Matrix Riccati Equations in Control and Systems Theory*. Basel: Birkhäuser, 2003.
14. Bertsekas DP and Tsitsiklis JN. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
15. Werbos PJ. *Beyond Regression: New Tools for Prediction and Analysis in the Behavior Sciences*, PhD Thesis, 1974.
16. Werbos PJ. Approximate dynamic programming for real-time control and neural modeling. In White DA and Sofge DA (eds), *Handbook of Intelligent Control*. New York: Van Nostrand Reinhold, 1992.
17. Haywood OG. Military decision and game theory. *Journal of the Operations Research Society* 1954; 2: 365–385.
18. Cantwell GL. *Can Two Person Zero Sum Game Theory Improve Military Decision-Making Course of Action Selection*? Monograph. Fort Leavenworth, KS: US Army School of Advanced Military Studies, 2003.
19. Mockaitis T. *Civil–Military Cooperation in Peace Operations: The Case of Kosovo*. Monograph. Carlisle, PA: United States Army War College, Strategic Studies Institute, 2004.
20. Vamvoudakis KG and Lewis FL. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica* 2010; 46: 878–888.
21. Vamvoudakis KG and Lewis FL. Online actor critic algorithm to solve the continuous-time infinite horizon optimal control problem. In *Proceedings of the International Joint Conference on Neural Networks*, Atlanta, GA, 2009, pp. 3180–3187.

Authors’ biographies

Greg Hudas is the Acting Director, Joint Center for Robotics (JCR) at the U.S. Army Tank–Automotive Research, Development, and Engineering Center (RDECOM/TARDEC), Warren, Michigan. One of his major responsibilities includes directing the Ground Robotics Reliability Center comprised of universities, small companies, and traditional defense OEMs in support of U.S. Army Ground Robotics Objectives. Other duties include providing S&T guidance to the Robotics Systems Joint Project Office (RS-JPO) and the OSD Joint Ground Robotics Enterprise (JGRE). He holds BS, MS, and PhD degrees in Electrical and Computer Engineering with research emphasis in intelligent systems and control of mobile robotic systems. He has authored/co-authored over 40 technical journal articles and major conference papers and is currently a visiting Research Scholar at the University of Michigan.

Kyriakos G Vamvoudakis was born in Athens, Greece. He received the Diploma in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2006 with highest honors and the MSc degree in Electrical Engineering from The University of Texas at Arlington in 2008. He is currently working toward the PhD degree and working as a research assistant at the Automation and Robotics Research Institute, The University of Texas at Arlington. His current research interests include approximate dynamic programming, neural network feedback control, optimal control, adaptive control and systems biology. He is a member of Tau Beta Pi, Eta Kappa Nu and Golden Key honor societies and is listed in *Who's Who in the world* and *Who's Who in science and Engineering*. He is a registered electrical/computer engineer (PE) and member of Technical Chamber of Greece. He has authored and coauthored over 15 technical publications.

Dariusz Mikulski is a PhD candidate in Electrical and Computer Engineering at Oakland University in Rochester Hills, Michigan. He earned his BSE in Computer Science from the University of Michigan in Ann Arbor, MI and MS in Computer Science and Engineering from Oakland University. His current research interests include: multi-agent systems, pattern recognition, and machine learning. He is also the University Relations Liaison, National Automotive Center (NAC) at the U.S. Army Tank–Automotive Research Development and Engineering Center (RDECOM/ TARDEC), Warren, Michigan. His primary duty is to develop collaborative relationships between Academia and TARDEC in order to advance the US Army's strategic research thrusts.

Frank L Lewis, Fellow IEEE, Fellow IFAC, Fellow U.K. Institute of Measurement & Control, PE Texas, U.K. Chartered Engineer, is Distinguished Scholar Professor and Moncrief-O'Donnell Chair at University of Texas at Arlington's Automation & Robotics Research Institute. He obtained the Bachelor's Degree in Physics/EE and the MSEE at Rice University, the MS in Aeronautical Engineering from University of West Florida, and the PhD at Ga. Tech. He works in feedback control, intelligent systems, distributed control systems, and sensor networks. He is author of 6 U.S. patents, 216 journal papers, 330 conference papers, 14 books, 44 chapters, and 11 journal special issues. He received the Fulbright Research Award, NSF Research Initiation Grant, ASEE *Terman Award*, International Neural Network Society *Gabor Award* 2009, and U.K. Institute of Measurement and Control *Honeywell Field Engineering Medal* 2009. He has received an Outstanding Service Award from Dallas IEEE Section, been selected as Engineer of the year by Ft. Worth IEEE Section, listed in Ft. Worth Business Press Top 200 Leaders in Manufacturing, received the 2010 IEEE Region 5 Outstanding Engineering Educator Award and the 2010 UTA Graduate Dean's Excellence in Doctoral Mentoring Award. He served on the NAE Committee on Space Stations in 1995. He is an elected Guest Consulting Professor at South China University of Technology and Shanghai Jiao Tong University. He is a founding Member of the Board of Governors of the Mediterranean Control Association. He helped win the IEEE Control Systems Society Best Chapter Award (as Founding Chairman of DFW Chapter), the National Sigma Xi Award for Outstanding Chapter (as President of UTA Chapter), and the US SBA Tibbets Award in 1996 (as Director of ARRI's SBIR Program).